BMJ

LONDON, SATURDAY 11 JUNE 1994

Increasing the accessibility of data

"See for yourself" should be the watchword

In 1849 the Philadelphia physician Samuel George Morton claimed that black people had smaller cranial capacities than white people.[1] As Morton included raw data in his publications Stephen Jay Gould was able to reanalyse them 130 years after Morton's death.[1,2] What Gould found was that there were no grounds for Morton's claims: the data had been manipulated to support the investigator's prior hypothesis.

By publishing his data in full Morton was unconsciously engaging in sharing data. While providing access to the full details and results of experiments has long been considered a characteristic ethos of science,[3] discussion of the principles and practicalities has been more recent.[2,4-7] Various benefits of sharing data can be identified.[2,7,8] Firstly, as the collection of data generally constitutes the main cost of studies, the use of existing data to answer issues not directly addressed by the primary researchers represents an efficient use of resources. Sharing data can also reduce the burden imposed upon study participants, both for individuals and groups who are at risk of becoming overresearched. Secondly, replicating the findings of one study within other datasets increases their robustness. Thirdly, when an investigation is being planned, analysis of data from earlier studies can help in the formulation of the research question, the refinement of measurement instruments, and the calculation of sample sizes.

Fourthly, new datasets can be created through linkage of different sets of records on the same people. Fifthly, exercises that combine data, such as meta-analysis, build on results of primary studies. Published summary statistics have generally been the level at which data have been aggregated, but as meta-analyses combining data on individual patients become more common,[9,10] their dependence on formal data sharing will increase.

Finally, the ability of outside researchers to check whether the conclusions drawn from an analysis are justified increases confidence in these conclusions. An extreme case relates to the detection of fraud: if access to the original data from published studies was routine then faking results would become more tricky. Different approaches to data analysis, driven partly by differing prior opinions, however, is probably a more common cause of disagreement among researchers than straightforward dishonesty. Thus two coinvestigators of a randomised controlled trial of antimicrobial treatment for otitis media fundamentally disagreed over the analysis of the results and produced separate reports, coming to opposite conclusions.[11,12] Even results from randomised controlled trials are therefore not immune to the influence of the data analyst. With observational data the ability to reach different conclusions is more widely recognised.[13-15] In such cases it is often not that one analyst is right and the other wrong but that different assumptions, implemented through different analytical strategies, can produce conflicting results. Sharing data can ensure that when truth cannot be guaranteed it is not simply imposed by fiat.

Objections to data sharing from primary researchers are understandable. The hard work of running and reporting a study could be compounded by having to prepare your dataset for wider access, then being haunted by the possibility of your conclusions being rubbished on the basis of reanalysed data.[16] Ways to protect the interests of the primary researcher include guaranteeing the rights to initial publication, transferring the costs of preparing databases to secondary analysts, ensuring no commercial exploitation of shared data by the recipient, and constructing formal agreements on sharing data.[17] In this way one of the fears about sharing data—that the rewards for collecting data would be so reduced as to have a detrimental effect on the progress of science[16]—can be allayed. Similarly, the assessment of research output should take into account the high contribution of researchers who collect data that serve as the basis for secondary analyses.

The tradition of trade secrecy runs against the ethos of sharing data, and commercial objections to greater access to data may be raised.[18] Although corporations can legally gain access to data that they regard as contrary to their interests[4,19] —as happened during the litigation regarding alleged harmful effects of pertussis vaccine[20,21]—the situation is not reciprocal. Many unpublished data from sponsored pharmaceutical studies never become public, for the simple reason that it would be against the interests of the companies.[22] While several proposals exist for remedying this,[23] increasing the degree to which data sharing is seen as the expected norm, supported by professional associations and regulatory bodies, could be particularly important.[19]

What practical steps can be taken to encourage sharing of data? Grant giving bodies could make funding conditional on willingness to share data. In the United States this is increasingly practised by government agencies,[6] and in Britain the Economic and Social Research Council—but not the Medical Research Council—has adopted this practice. Plans for sharing data could become a requirement for protocols, upholding the reasonable view that data paid for by public money are public property. Planning in advance to share data would ensure that databases allow for the production

of datasets for secondary analysis that maintain the confidentiality of subjects. Adequate documentation should accompany such datasets, and the costs of producing the material should be covered by the initial grants—in which case funding agencies would have to provide the necessary finance—or costs should be covered by the secondary analysts.

Clearing houses for shared data would reduce the burden on primary researchers. After providing a documented dataset to the central archive once, researchers need not repeatedly answer requests, as the archive takes on this task. The Economic and Social Research Council has established such an archive,[24] and other topic based databanks exist.[6]

Journals also have a role in encouraging the sharing of data. For example, the *American Journal of Public Health* stipulates that data are, in principle, available to the editors and to interested researchers.[6] If papers submitted to journals had to be accompanied by a disk copy of the data on which they were based then statistical referees could check that the results were not the product of overenthusiastic data torture.[25] In this way sharing of data could contribute to improving the quality of published research. The fall in submissions to a journal brave enough to implement this policy would be a useful indicator of its success.

GEORGE DAVEY SMITH
Senior lecturer in epidemiology and public health

Department of Public Health,
University of Glasgow,
Glasgow G12 8RZ

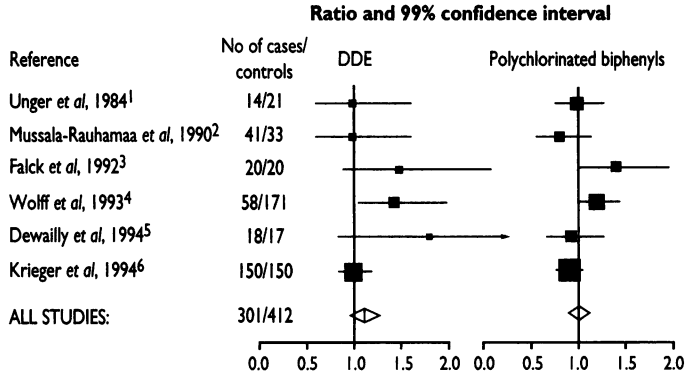
1 Gould SJ. *The mismeasure of man*. New York: WW Norton and Co, 1981.
2 Fienberg SE, Martin ME, Straf ML. *Sharing research data*. Washington DC: National Academy Press, 1985.
3 Metton RK. Science and the social order. *Philosophy of Science* 1938;5:321-37.
4 Yolles BJ, Connors JC, Grufferman S. Obtaining access to data from government-sponsored medical research. *N Engl J Med* 1986;315:1669-72.
5 Melton GB, ed. Must researchers share their data? *Law and Human Behavior* 1988;12:159-206.
6 Marshall E, Roberts L. Data sharing: a declining ethic? *Science* 1990;248:952-7.
7 Hogue CJR. Ethical issues in sharing epidemiologic data. *J Clin Epidemiol* 1991;44(suppl 1): 103-7S.
8 Hedrick TE. Justifications for the sharing of social science data. *Law and Human Behavior* 1988;12:163-71.
9 Pignon J-P, Arriagada R, Ihde DC, Johnson DH, Perry BC, Souhami RL, *et al.* A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *N Engl J Med* 1992;327:1618-24.
10 Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* 1993;341:418-22.
11 Cantekin EZ, McGuire TW, Griffith TL. Antimicrobial therapy for otitis media with effusion (secretory' otitis media). *JAMA* 1991;266:3309-17.
12 Mandel EM, Rockette HE, Bluestone CD, Paradise JL, Nozza RJ. Efficacy of amoxycillin with and without decongestant - antihistamine for otitis media in children. Results of a double-blind randomized trial. *N Engl J Med* 1987;316:432-7.
13 Johnson CL, Woteki CE. The art and science of interpreting survey data. *Am J Public Health* 1990;80:1427-9.
14 Subjectivity in data analysis [editorial]. *Lancet* 1991;337:401-2.
15 Phillips AN, Davey Smith G. How independent are "independent" effects? Relative risk estimation when correlated exposures are measured imprecisely. *J Clin Epidemiol* 1991;44:1223-31.
16 Stanley B, Stanley M. Data sharing: the primary researcher's perspective. *Law and Human Behavior* 1988;12:173-80.
17 Sieber JE. Data sharing: defining problems and seeking solutions. *Law and Human Behavior* 1988;12:199-206.
18 Fayerweather WE, Tirey SL, Baldwin JK, Hoover BK. Issues in data sharing and access: an industry perspective. *J Occup Med* 1991;33:1253-6.
19 Cecil JS, Boruch R. Compelled disclosure of research data. *Law and Human Behavior* 1988;12: 181-9.
20 Miller DL, Wadsworth MJH, Ross EM. Pertussis vaccine and severe acute neurological illnesses. *Vaccine* 1989;7:487-9.
21 Bowie C. Lessons from the pertussis vaccine court trial. *Lancet* 1990;335:397-9.
22 Mindel JS. Failure of controlled clinical trial data to reach the literature. *Clin Pharmacol Ther* 1992;52:4-5.
23 Levy G. Publication bias: its implications for clinical pharmacology. *Clin Pharmacol Ther* 1992;52:115-9.
24 The ESRC Data Archive. *Sharing and preserving research data*. Colchester: ESRC Data Archive, 1993.
25 Altman DG. The scandal of poor medical research. *BMJ* 1994;308:283-4.

Organochlorines in the environment and breast cancer

The data so far produced provide reassurance rather than anxiety

Women have been worried recently by press accounts of six comparatively small epidemiological studies suggesting that certain organochlorines in the environment might increase the risk of breast cancer.[1-6] The studies were concerned with 1,1-dichloro-2,2-bis(*p*-chlorophenyl) ethylene (DDE), the main metabolite of the insecticide DDT and also with the polychlorinated biphenyls (PCBs), a group of compounds used in industry—for example, as electrical insulators. How strong are the grounds for concern?

The results are summarised in the figure. The individual estimates plotted are ratios of the mean concentrations of DDE or of polychlorinated biphenyls in fat or serum in women with breast cancer divided by the mean concentrations in control women without breast cancer. Summary ratios were derived from the weighted averages of the log ratios. For



Ratios of mean concentrations of DDE and polychlorinated biphenyls in cases and controls from six studies. Areas of solid squares are proportional to information in each study

DDE the summary ratio was 1·11 (99% confidence interval 0·97 to 1·26). In other words, the women with breast cancer had slightly higher concentrations of DDE than controls but the difference was not statistically significant. For the polychlorinated biphenyls the summary ratio was 1·01 (0·92 to 1·10)—indistinguishable from no difference at all.

When they were first published two of the studies—by Falck and colleagues[3] and by Wolff and colleagues[4]—created considerable anxiety with respect to DDE, but they were based on only 20 and 58 women with breast cancer respectively. The recent study of Krieger and colleagues[6] had as many cases as all the previous studies put together and used serum collected an average of 14 years before diagnosis, whereas the other studies used fat or serum collected shortly before or after diagnosis. Krieger and colleagues found no association of DDE with breast cancer.

Widespread use of DDT began in the United States in 1946 and increased until 1959. It then declined steadily until it effectively stopped in 1972.[7] DDT accumulates in the body, but the reduction in its use caused concentrations in adipose tissue in the general population to fall from about 8 ppm in 1970 to about 2 ppm by 1983.[7] Polychlorinated biphenyls were first produced commercially in 1929 and since then have been used in many industrial products. They were detected as environmental contaminants in 1966, and in the United States production ceased in 1977. In 1972, 61% of the American population was estimated to have concentrations of polychlorinated biphenyl in adipose tissue above 1 ppm, but this had dropped to 6% by 1983.[7] In most developed countries patterns of use of both types of organochlorines have resembled those in the United States, but DDT